A TESOL Symposium

English Language Assessment

Kyiv National Taras Shevchenko University
Kyiv, Ukraine
Friday, October 26, 2007

Featured Speakers

Jeffrey T. Connor-Linton
Constant Leung
Jo Lewkowicz

Closing Session Facilitator:

Olga Bessonova

TESOL Symposium Sponsors:

<table>
<thead>
<tr>
<th>2007 TESOL Symposia Global Sponsor</th>
<th>TESOL Symposium Sponsor</th>
</tr>
</thead>
<tbody>
<tr>
<td>HEINLE CENGAGE Learning</td>
<td>The Seifert Fund</td>
</tr>
</tbody>
</table>

TESOL acknowledges and appreciates the assistance of TESOL Ukraine and the U.S. Department of State, Office of English Language Programs.
Contents

- INTRODUCTION
  - Overview
  - Symposium Agenda
  - Speakers Bios and an Overview of the Presentations

- PAPERS OF THE PRESENTERS
  - The Importance of Being Transparent: Benefits to Society, Schools, and Students of Using a Common Language Assessment Framework (Pages 1-10)
    Jeff Connor-Linton
  - Formative Assessment: Time of Learning (Pages 1-10)
    Constant Leung
  - Should Authenticity Continue to Trouble Language Testers (and Teachers)? (Pages 1-8)
    Jo Lewkowicz
  - Closing Session: Summary (Pages 1-8)
    Olga Bessonova
TESOL Symposium on English Language Assessment

Overview

On October 26, 2007, 199 participants gathered at Kyiv National Taras Shevchenko University, in Kyiv, Ukraine, for the TESOL Symposium on English Language Assessment. The featured speakers were Jeffrey T. Connor-Linton, Georgetown University, Washington, D.C., United States, Constant Leung, King’s College London, London, England, and Jo Lewkowicz, University of Warsaw, Warsaw, Poland. Olga Bessonova, Donetsk National University, Donetsk, Ukraine, led the closing session. TESOL Ukraine, TESOL’s affiliate, chose the theme of the symposium.

Sandy Briggs, President of TESOL, welcomed the participants and gave special thanks to the symposium sponsors: Heinle, a part of Cengage Learning, and The Seifert Fund. She also acknowledged and thanked TESOL Ukraine and Kyiv National Taras Shevchenko University. Svetlana Radzievskaya, President, TESOL Ukraine, and Petro Bekh, Vice Rector for Foreign Relations, Kyiv National Taras Shevchenko University, also provided opening remarks.

In the morning session, Connor-Linton, Leung, and Lewkowicz shared highlights of their work and research. In the afternoon, each of the speakers led an interactive breakout session. In the wrap-up session, Olga Bessonova summarized key insights from the symposium presentations and papers and facilitated a discussion involving the speakers and participants.

The three original papers and summary of the proceedings published here provide pertinent research and insights on the important topic of English language assessment.
TESOL Symposium on English Language Assessment

Symposium Agenda

October 26, 2007

Kyiv National Taras Shevchenko University

Kyiv, Ukraine

8:30 am—9:30 am  Check in/Tea and Coffee

9:30 am—9:45 am  Welcome

- Sandy Briggs, President, Teachers of English to Speakers of Other Languages, Inc.
- Svetlana Radzievskaya, President, TESOL Ukraine
- Petro Bekh, Vice Rector for Foreign Relations, Kyiv National Taras Shevchenko University

9:45 am—12:00 pm  Introductory Presentations

- The Importance of Being Transparent
  Jeffrey T. Connor-Linton
- Formative Assessment: Time of Learning?
  Constant Leung
- Should Authenticity Continue to Trouble Language Testers (and Teachers)?
  Jo Lewkowicz

12:00 pm—1:30 pm  Lunch

1:30 pm—3:15 pm  Concurrent Discussions

- The Importance of Being Transparent
  Jeffrey T. Connor-Linton
- Formative Assessment: Time of Learning?
  Constant Leung
- Should Authenticity Continue to Trouble Language Testers (and Teachers)?
  Jo Lewkowicz

3:30 pm—4:30 pm  Closing Session/Questions & Comments

Olga Bessonova
TESOL Symposium on English Language Assessment

Speaker Bios and Overview of the Presentations

The Importance of Being Transparent: Benefits to Society, Schools, and Students of Using a Common Language Assessment Framework

An important dimension in the development of language testing is the improvement of transparency—increasing explicitness about what is being tested, how it is being tested, and how test results are used. This paper discusses examples of how transparency benefits language testing in international, societal, institutional, and pedagogical contexts. At the international level, shared definitions of language use and desired learning outcomes can facilitate comparability and cooperation among educational institutions in different countries. At the societal level, more explicit language learning standards and testing methods can improve the fairness and face validity of tests and their use in providing access to social resources. At the institutional level, explicit descriptions of desired learning outcomes can enhance articulation of levels of instruction and encourage positive washback from assessment to instruction. At the pedagogical level, sharing evaluation rubrics with language learners links summative and formative assessment, helping them assess their own language learning and use outside the classroom.

Jeff Connor-Linton is an associate professor of linguistics at Georgetown University, Washington, D.C., United States, where he has been department chair, head of applied linguistics, and director of the MA-TESL program. He uses multidimensional analysis of discourse to investigate processes of cross-cultural (mis)communication, language assessment, and second language acquisition and writing. In addition to publishing his research, Jeff has consulted on curriculum and assessment for public school foreign language programs and for IEPs worldwide. He is currently co-director of a study of the effects of studying abroad on learners’ oral proficiency and intercultural development. He is also developing an online database of second language learner performance data. A member of the American Association for Applied Linguistics (AAAL) since 1991, Jeff was recently elected second vice-president. He will be program chair for the 2009 AAAL Annual Conference in Denver before becoming president.

Formative Assessment: Time of Learning?

In the past ten years or so, interest has intensified in formative assessment that foregrounds learning, often referred to as assessment for learning (AfL) in the current educational assessment literature. It has gained considerable recognition by teachers and researchers working in different curriculum subjects because this form of assessment claims to promote student learning. Formative assessment has also begun to achieve a high degree of influence in the educational policies of Canada, Hong Kong, Scotland, and Wales. This paper examines the key principles of formative assessment with particular reference to second/foreign language education. Classroom data is used to illustrate this form of assessment in action. Conceptual issues such as formative direction and professional issues such as teacher preparation are also addressed.

Constant Leung is a professor of educational linguistics and deputy head of department in the Department of Education and Professional Studies, King’s College London, United Kingdom. Before taking up teaching positions in higher education he taught in schools and worked as advisory
teacher and manager in local government. He is director of two programmes at King’s College London: MA English Language Teaching and Applied Linguistics, and Academic Language Development. He was the founding chair of the National Association for Language Development in the Curriculum, United Kingdom. Currently he serves as chair of TESOL’s Standing Committee on Research; he also serves on the editorial boards of four international research journals. His research interests include education in ethnically and linguistically diverse societies, second/additional language curriculum development, language assessment, language policy, and teacher professional development. He has written and published widely on issues related to ethnic minority education, additional/second language curriculum, and language assessment nationally and internationally.

**Should Authenticity Continue to Trouble Language Testers (and Teachers)?**

The issue of authenticity in language testing has been extensively discussed and our understanding of this notion has developed considerably. Yet, to what extent has the debate facilitated test development? This paper reviews the notion of test authenticity and looks at some of the outstanding issues in relation to developing both high stakes tests as well as lower stakes, teacher-developed classroom assessments.

**Jo Lewkowicz** is a professor of applied linguistics at the University of Warsaw, Poland. Prior to her current appointment in Poland, she taught at tertiary level in Armenia, Hong Kong, China, Kenya, and Egypt. Her main areas of teaching, research interest, and publications are various aspects of language testing and assessment as well as academic literacies.

**Closing Session**

**Olga L. Bessonova** graduated from Moscow State Linguistic University, Russia, and has served as the vice-rector for international relations at Mariupol Institute of Humanities. She is currently head of the English Philology Department and Doctor of Philology, Professor at Donetsk National University, in Donetsk, Ukraine.
The Importance of Being Transparent: Benefits to Society, Schools, and Students of Using a Common Language Assessment Framework

Jeff Connor-Linton
Georgetown University
Washington, D.C., United States

In Oscar Wilde’s play, ‘The Importance of Being Earnest,’ being earnest has many meanings. Similarly, transparency in language testing has many values, at different levels of application. The goal of this paper is to make the case for bringing transparency to language testing by briefly describing and exemplifying some of its benefits—at the international, national, institutional, and classroom levels. A transparent object is, of course, one that can be seen through. In governmental and institutional contexts, transparency implies openness and accountability; when evaluative processes and standards can be reviewed by anyone, there is less opportunity for abuse of the system and stronger belief among stakeholders in the fairness of the evaluation system. In pedagogical contexts, transparency not only enhances the face validity of a test, it allows learners to internalize the standards against which their performances are judged. A transparent test creates beneficial washback, which is when a test motivates positive teaching and learning behaviors and thereby becomes a valuable pedagogical tool, not just an evaluative weapon.

Transparency at the International Level

One way in which language testers can seek to attain transparency is by adopting shared, explicit standards of language performance which provide a common basis for comparison. In a sense, test results are a currency—a currency that provides access to social capital, like education and jobs. The more places a currency is accepted, the more stable and valuable that currency is. By developing transparent testing practices, an educational system allows the marketplace to more easily understand the value of that system’s products—its graduates. If the marketplace cannot easily understand a product, it won’t value that product—or the educational system that produced it.

Hansche (1999) defines performance standards as a system including performance levels, descriptors of performance at each level, exemplars of student work at each level, and cut-off scores that separate the adjacent levels of performance. This simple definition belies the amount of cost and effort it takes to build such a system, and suggests one of the main benefits to a country of sharing a common set of performance standards with other countries: The cost of developing an internally coherent set of performance standards, measuring its reliability, and mounting a convincing argument for its validity is very high. If a country can adopt or adapt an existing set of performance standards, it avoids reinventing an entire language assessment system, and it can use its limited language education resources more efficiently. What a country ‘loses’ in autonomy, it more than gains in efficiency.

There are a variety of more specific benefits to a country of being transparent by adopting shared performance standards. Ideally, such standards help members of a community become more efficient in their language teaching and testing. When language tests are benchmarked to
common standards, they can be more easily adapted to new languages, and to new populations of language learners and new domains of language use within a given target language. Language professionals can develop tests more efficiently within a common framework, and interpret the results of each others’ tests more easily. Being explicitly, transparently related to a set of common performance standards enhances the value of a test, because users—employers, other academic institutions, teachers, and students—share a common understanding of what a particular rating on the test means. Employers can relate language requirements for particular jobs to level descriptors of the common standards, and be confident of the language skills of those they hire. Schools can rely on ratings of students’ language abilities and place them more accurately in the proper curricula (saving educational costs and improving student motivation and performance). Teachers can rely on the language ratings of new students in their classes and more efficiently respond to those students’ individual needs. And students can use the test’s performance descriptors as a rubric for self-assessment and a tool for more effective language learning inside and beyond the classroom.

The Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR; Council of Europe; n.d.) provides just such a set of shared performance standards. It “describes in a comprehensive way what language learners have to learn to do in order to use a language for communication and what knowledge and skills they have to develop so as to be able to act effectively” (Council of Europe, n.d., p.1). The CEFR describes (i) the competences necessary for communication, (ii) related knowledge and skills, and (iii) the situations and domains of communication. It helps language teaching professionals to set standards of performance to be attained at successive stages of learning and to evaluate test performances in an internationally comparable manner. The CEFR provides a basis for the mutual recognition of language qualifications, thus facilitating educational and occupational mobility.

The Common European Framework of Reference for Languages

Although some readers of this paper may be familiar with the CEFR (Council of Europe, n.d.), a short summary of a few of its main features will serve as a foundation for the arguments that follow. Based on extensive prior research and current theories of language acquisition, the CEFR posits three main stages of development, each with two sublevels, as shown in Figure 1. Each sublevel bears a general descriptive name, carried over from previous research.

![Figure 1. Stages of Development (Council of Europe, n.d., p. 23)](attachment:figure1.png)
Each sublevel (A1 through C2) is described in several ways, and the intersection of these different foci is a valuable resource for development of curricula and learning activities, as discussed below. Each sublevel is described globally, in sets of skill-related ‘can-do’ statements. These statements may be used for self-assessment by learners (as in the example below) or for other-assessment by teachers and testers.

For example, according to the CEFR (Council of Europe, n.d.), a Basic User at the A1 stage

Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help. (p. 26)

In comparison, an Independent User at the B1 stage

Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans. (p. 24)

Each level subsumes the levels below it on the scale; so, for example, someone at the B1 level of development is considered also to be able to do what is stated at A2 and A1.

These global ‘can do’ descriptors are further specified in relation to the skills of listening, reading, writing, and speaking (both interaction and monologic production). So, for example, the A1 Basic User in listening “can recognize familiar words and very basic phrases concerning [herself], [her] family and immediate concrete surroundings when people speak slowly and carefully” (Council of Europe, n.d., p. 26), whereas the B1 Independent User “can understand the main points of clear standard speech on familiar matters regularly encountered in work, school, leisure, etc. . . . [and] can understand the main point of many radio or TV programmes . . . when the delivery is relatively slow and clear” (Council of Europe, n.d., p. 26). There are similar descriptors for reading, spoken interaction, spoken production, and writing.

The levels can be further specified—for example, in terms of “qualitative aspects of spoken language use,” (Council of Europe, n.d., p. 25) like range, accuracy, fluency, interaction, and coherence. For example, the A1 Basic User has a limited lexicon and controls just a few formulaic grammatical structures which force her to rely on strategic competence and the kindness of her interlocutors to accomplish even simple interactions. In comparison, the B1 Independent User’s vocabulary extends to more domains and more complex grammatical patterns stitched together into templates of conversational routines, with less reliance on her interlocutor to construct longer and more complex discourse structures.
The CEFR (Council of Europe, n.d.) also suggests the genres, domains, and topics that can be handled by learners at each level of development. For example, within the genre of Describing and Narrating, the A1 Basic User can talk about where she lives, while the B1 Independent User can tell a story and give her reaction to it.

Even this brief and simplistic overview of the investment that has been made in the development of the CEFR (Council of Europe, n.d.)—and the resource that it offers to countries working to develop their own language curricula and assessment systems—strongly suggests that there are tremendous benefits of membership (a good capitalistic advertising phrase!) for the larger European language teaching and testing community.

Transparency at the National Level

So far, I have discussed the value of transparency in language testing mainly in international terms. Within a country, transparency in language testing helps promote equal access to educational resources and thereby enhances the reputation of the educational system and the government that sponsors that educational system.

Ukraine’s *National Doctrine of Education Development in the 21st Century* (Second National Workers Congress, 2001), adopted in 2002, recognizes education as a priority of its national policy, and in particular prioritizes the promotion of equal access to quality education. In April 2006, the Ministry of Education and Science of Ukraine initiated a 10-year education reform program, with support from the World Bank, for the implementation of the Equal Access to Quality Education Project. A critical element of equality of access to education is fairness in testing.

Adoption or adaptation of explicit, detailed and transparent descriptions of the stages of second language development go a long way toward meeting the standards of fair testing, like the Code of Fair Testing in Education (2004) developed by the Joint Committee on Testing Practices of the American Psychological Association. Transparency of the criteria of evaluation helps ensure accountability of test developers and administrators and therefore the fairness of tests.

An explicit framework of reference like the CEFR (Council of Europe, n.d.) allows test developers, according to the Code of Fair Testing Practices in Education (2004), to “provide the information and supporting evidence that test users need to select appropriate tests” and allows test users to “define the purpose for testing, [and] the content and skills to be tested” for a group of test takers. Such a framework helps test developers to “explain how to administer and score tests correctly and fairly” and provides the necessary basis for test administrators to “provide adequate training to scorers and ensure and monitor the accuracy of the scoring process.” Common reference levels of language development “provide information to help test users interpret test results correctly.” The multiple perspectives on learners’ language performances ensure that “test users . . . base decisions about test takers on multiple sources of appropriate information, not on a single test score.” Finally, a transparent framework for language assessment provides a concrete reference for “inform[ing] test takers . . . about the coverage of the test, the types of question formats, the directions, and appropriate test-taking strategies.”
When evaluative standards can be reviewed by anyone, there is less opportunity for abuse of the system and stronger belief among stakeholders in the fairness of the evaluation system—and the institution doing the evaluating. Conversely, without transparent evaluative criteria to mediate between the learner’s (often inflated) self-perception and the tester’s assessment, the fairness of even the best-intentioned, most scrupulous evaluator frequently will be doubted, precisely because there is no independent framework for comparison to refer to.

It is worthwhile to note that in addition to the extensive documentation of the CEFR (Council of Europe, n.d.) itself, which is available on the Internet, there is also a preliminary pilot version of a Manual for Relating Language Examinations to the Common European Framework of Reference of Languages (Council of Europe, 2003). This manual is offered to help test developers to relate their examination(s) to the CEFR, first by describing the coverage of a test, how it is administered, and how its results are analyzed, and then by relating those results to the common reference levels of the CEFR. The manual does not prescribe any one approach to constructing language tests or examinations; it is designed to allow test developers to relate tests developed according to a range of testing approaches to a common set of performance standards.

Transparency at the Institutional Level

In the development of the CEFR (Council of Europe, n.d.), great care was taken to ensure that teachers could relate to the descriptive categories selected, and that descriptors actually described the categories they were intended to describe. Finally, the best descriptors in the set were scaled using quantitative methods. The accuracy of this scaling has since been checked in replication studies. (p. 22)

Language use and language learning are complex systems, and characterizing levels of development from several perspectives enriches the description by allowing the teacher or tester to triangulate the different perspectives in order to reach a more reliable and useful assessment of the individual learner. In addition to providing descriptors that allow a teacher or tester to identify a learner’s level of language development, the multiperspective approach of the CEFR (Council of Europe, n.d.) is also a rich resource for curriculum development.

A framework like the CEFR (Council of Europe, n.d.) is not just for test developers; it defines teaching and learning objectives and methods for course designers, textbook writers, teachers, and teacher trainers. Tyler (1950) identified the necessary elements of curriculum design as 1) the educational purposes that the institution seeks to attain; 2) the educational experiences provided to attain these purposes; 3) the organization of these educational experiences; and 4) the evaluation of these experiences. A framework like the CEFR offers a good deal of guidance in developing or redeveloping a foreign language curriculum by providing “a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc. . . .” (Council of Europe, p. 1). Because it is relevant and accessible to all stakeholders in language education within a society, it can support both top-down and bottom-up processes of curriculum reform (e.g., Connor-Linton, 1996).
As described in my account of a foreign language curriculum redevelopment process in the public schools of Arlington, Virginia, USA, a curriculum is successful only if it is a valuable resource and guide for teachers’ syllabus design and class planning (Connor-Linton 1996). A curriculum which does not make teachers’ planning work easier will fall into disuse, and articulation of instruction across levels of the curriculum will suffer. One way to make a curriculum useful to teachers is to structure it like a restaurant’s menu of level-ordered options, from which a teacher may choose in planning pedagogical activities. (In some restaurants in the United States, menus are organized into columns or sections, and a fixed price meal is built by ordering one item from Column A, one item from Column B and so on.) Recall that the CEFR (Council of Europe, n.d.) describes i) the competences necessary for communication, ii) the related knowledge and skills, and iii) the situations and domains of communication—at each level of development. So, using a framework like the CEFR as the basis for their curriculum, teachers can combine these elements from a target level to plan a class or sequence of classes that integrates the skills within a particular domain of language use, or allows students to practice a given skill across several domains.

For example, imagine that a teacher of a class of B1 Independent Users wants to plan a sequence of classes. From just the brief subset of CEFR (Council of Europe, n.d.) can-do descriptors mentioned earlier, the teacher can select a domain, like “radio or TV programmes . . . [in which] the delivery is relatively slow and clear,” select a relevant skill (listening), and select a genre (narration); this set of intersecting features from the framework not only suggests a range of activities, but also indicates the target of instruction and evaluation (main points of the program) and provides the standards for student assessment. The teacher can then focus on a different skill (speaking), but maintain the genre (narration) by asking students to retell the plot of the program they have just heard. They can then assess the range of the students’ linguistic forms, the accuracy of use of those forms, and the students’ fluency and coherence, in relation to the qualitative descriptors of speaking at the B1 level of the CEFR. And then the teacher might ask the students to write an essay describing their impressions of the program they’ve seen, and have them read and critique each other’s essays, using the B1 level descriptors for writing. For another sequence of classes, the teacher might ask students to practice speaking in a range of situations that might arise while traveling.

The many intersections of the different aspects of the framework at a given level of proficiency offer a rich resource for planning pedagogical activities as well as providing the criteria against which student language use in those activities may be evaluated.

**Transparency at the Classroom Level**

A transparent framework for assessment, like the CEFR (Council of Europe, n.d.), is a rich resource for language learners, as well as teachers. I will briefly describe a strategy through which a common framework of language development can be used to help students to internalize the goals of instruction and apply them to their own language learning. The general strategy is to work with students to develop explicit rating rubrics based on the broader reference levels of the framework.
First, consider the options for rating student writing in a second language writing course. To take an extreme example, if the teacher simply rates the papers, on some general scale (A, B, C, . . . or 5, 4, 3, 2, 1) without providing further feedback, the student has received a general indicator of the quality of her writing, relative to some seemingly secret scale of value. The diagnostic value of this feedback is limited, and disproportionate to the time and effort put into the paper by both the student and the teacher. These grades provide little diagnostic information; even students who receive high ratings do not know what, specifically, they did well, and may inappropriately generalize strategies from a successful essay to a subsequent assignment. Such a nontransparent grading system is likely to be seen as arbitrary and unfair, particularly by students who receive low ratings. It concentrates authority in the hands of the teacher and deters students from taking responsibility for improving their writing.

To improve on this scenario, a teacher also may write marginal and/or concluding comments on the student’s essay. These comments go some way toward explaining the grade that the teacher has assigned, but often these comments focus students on correction of errors without providing direction for attaining higher levels of proficiency. More troubling, perhaps, is the large body of research that suggests that students’ writing is not substantively influenced by marginal comments.

Now, consider a writing class in which students use a rating rubric of explicit descriptors to rate and improve their own writing. How might this work? In broad strokes, first the teacher assigns a writing task, referring to the larger curricular framework, which is appropriate for the students’ level of proficiency. Then, students read and discuss a rating rubric based on the same framework. Next, students use the rubric to evaluate and provide feedback on their own and each other’s essays. Finally, students revise their own essays, turning in a record of the improvements they make and their reasons for making them (referring to the rating rubric) along with both drafts of their essays.

Let me elaborate a bit on this process. First, to be most useful, the rating rubric would include several levels of descriptors—below, at, and above the level of the students in the class (the entire range of the overall curriculum framework does not need to be used). Depending on the purposes and preferences of the teacher, the rubric might be holistic or analytic; that is, each level might be described by one general set of descriptors (statements about the writing produced at each level of development) or by several sets of descriptors, each of which describes different aspects of writing at each level of development (for example, grammar, vocabulary, coherence, and discourse structure). In either case, the descriptors should be as concrete as possible. The students should read and discuss the rubric. To contextualize that discussion and model successful writing, the teacher could provide one or two model essays, and students could discuss how the upper levels of the rating rubric apply to those model essays.

The students then apply the rating rubric to a set of essays which the teacher has previewed and rated; these essays should describe most of the range of writing ability in the class (lowest to highest). The first time a class tries this approach, these essays should be anonymous—for example, from another class or a previous cohort. Evaluating other, anonymous students’ writing facilitates more objective and critical evaluation, and students with lower writing skills are not embarrassed. It is also best if these anonymous essays are on the same topic as the current
students have written about; this will make it easier for students to make connections between the essays they are reading and improvements they can make on their own essays.

Independently, students read these essays, rate them using the rubric, and write down reasons for their ratings. This socializes students to acting as substantive editors. An ensuing class discussion can focus first on essays on which most students agreed in their ratings, then compare reasons for ratings of essays on which there was more disagreement. This discussion allows students to translate the terminology of the rubric’s level descriptors into their own terms and also can generate new, student-motivated level descriptors for the rubric.

At this point, each student applies the rating rubric to the (previously written) first draft of his or her own essay. (The teacher may also want to rate students’ first drafts so that students can compare their self-assessments to their teachers’ assessments.) Each student then revises his or her essay, keeping a record of the changes he or she has made and why he or she made them, referring to the descriptors on the rating rubric. The explicitness of this revision process helps students internalize the standards expressed in the rating rubric; they learn to use the rubric as a concrete guide for improvement of their own language use. The teacher can evaluate the revised essay, improvement from first to revised draft, and the quality of the revision process itself.

Getting students to evaluate their own language use works most easily with writing, but it can be adapted to speaking, by having students tape themselves and then rate their performances. It can also be adapted to reading and listening, for example, allowing students to rate language used by native speakers to understand more explicitly the models they are presented with. Involving students in evaluation helps them discover for themselves the concrete differences between less and more proficient language use. It helps them develop strategies for improving their own communicative skills, and gives them more control over their own second language development. Active involvement is a far more effective learning strategy than passive response to a teacher’s comments on a draft of an essay (e.g., Krashen, 1984). In an earlier study of this strategy (Connor-Linton, 1990), I noted several other pedagogical benefits of making evaluation transparent to learners:

[A] learner discovers what is important to her at that point in her individual development. The focus is shifted from the teacher and the teacher’s way of seeing writing to the students and their ways of seeing writing. The shift in focus allows student writers to look at their own writing critically and gives them some concrete tools with which to be their own editors, which can be adapted to a wide range of communicative situations and needs. (p. 354)

**Conclusion**

These examples have shown ways at each level of analysis—in international, national, institutional, and pedagogical—that transparency can enhance the value of language assessment. At the international level, it enhances and broadens the economic value of language ratings. At the national level, transparency of the criteria of language evaluation helps ensure accountability of test developers and administrators and therefore fairness of tests. At the institutional level, use of a transparent language assessment framework can facilitate language curriculum development.
and help make language curricula a vibrant resource for language teachers. Perhaps most important, at the classroom level, making language evaluation criteria transparent to learners can help them participate more actively in their own learning.

Transparency is the antithesis of autocracy. Instead of being enforced top-down, in transparent systems, standards are negotiated and maintained by sharing knowledge and authority. As history has taught us, concentration of power in the hands of a few is inefficient, at best, and this is as true in language testing as it is in politics more generally.
References


Formative Assessment: Time of Learning?

Constant Leung
King’s College London
London, England

Assessment has played a large part in education initiatives and reforms designed to improve student attainment in countries such as Australia, the United States and the United Kingdom in the past 15 years. In this talk, my focus is on (a) the pedagogic usefulness of classroom-based formative assessment in promoting learning and educational achievement and (b) some of the theoretical and practice-related issues that require further elaboration.

Teachers and Teaching Are Important

I will first use recent experience in England as a backdrop to provide a context for this discussion. In 1997 the government set an attainment target for 11-year-olds in state-funded schools—80% of the students in this age group were expected to achieve Level 4 in English by 2002 (DfEE, 1997). (Different targets were set for different school subjects.) The attainment levels are set out in the National Curriculum for the subject English (DfEE, 1997). Regular (formal standardized) testing of student performance was used as a policy instrument to push up achievement.

<table>
<thead>
<tr>
<th>Year</th>
<th>% pupils achieving level 4 or above in English</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996</td>
<td>57</td>
</tr>
<tr>
<td>1997</td>
<td>63</td>
</tr>
<tr>
<td>1998</td>
<td>64</td>
</tr>
<tr>
<td>1999</td>
<td>70</td>
</tr>
<tr>
<td>2000</td>
<td>75</td>
</tr>
<tr>
<td>2001</td>
<td>75</td>
</tr>
<tr>
<td>2002</td>
<td>75 Official target - 80%</td>
</tr>
<tr>
<td>2003</td>
<td>75</td>
</tr>
<tr>
<td>2004</td>
<td>78</td>
</tr>
<tr>
<td>2005</td>
<td>79</td>
</tr>
<tr>
<td>2006</td>
<td>79 Official target - 85%</td>
</tr>
<tr>
<td>2007</td>
<td>80</td>
</tr>
</tbody>
</table>

Note. Data is from Department for Children, Schools and Families (2007).

The published figures clearly show that there have been gains in the percentage of students achieving Level 4 in the past 10 years. But the official targets have not been met. However, throughout this period there have also been persistent reports that the high stakes nature of formal standardized testing has led to a narrowing of the English curriculum and teaching to the test. Press reports, for example, regularly suggest that in some primary schools the summer term has been given over to test preparation of various kinds. (See Leung & Rea-Dickins, 2007, for a further discussion.)
Although the British government has maintained its commitments to raising student achievements across the curriculum subjects, recently it has started actively encouraging teachers to adopt formative assessment, also known as assessment for learning (AfL) in their routine classroom practice (e.g., Qualifications and Curriculum Authority, n.d.). A strong argument for the promotion of AfL is that it can lead to enhanced student learning and achievement (Black & Wiliam, 1998). The Assessment Reform Group (2002) sets forth 10 principles of classroom practice that include the following statement:

Assessment for learning should be part of effective planning of teaching and learning. A teacher’s planning should provide opportunities for both learner and teacher to obtain and use information about progress towards learning goals. . . . Much of what teachers and learners do in classrooms can be described as assessment. That is, tasks and questions prompt learners to demonstrate their knowledge, understanding, and skills. What learners say and do is then observed and interpreted, and judgements are made about how learning can be improved. These assessment processes are an essential part of everyday classroom practice and involve both teachers and learners in reflection, dialogue, and decision making. (p. 2)

In general terms, AfL proponents advocate a range of teacher actions that includes

- using exploratory questioning and increasing wait time for student answers to generate opportunities for sustained discussion
- providing helpful feedback designed to improve student learning and performance
- using the information produced by summative tests for formative purposes.

These teacher actions, which can be understood in a commonsense way, can be adopted by teachers of all subjects. However for subject specialists to carry out AfL effectively, it is important for them to have an informed understanding of the key notions, such as exploratory questioning and helpful feedback, with reference to their subject discipline. For English language teachers, I suggest that the following issues are important:

- articulation and clarification of disciplinary belief and classroom practice at individual and collective levels (Ivanic, 2004; Leung 2005, 2007)
- attention to the co-constructed nature of classroom interaction and its implication for AfL in the English classroom (Leung, 2004; Leung & Mohan, 2004; McNamara 1997; Mohan, 1998; van Lier, 1996)
- professional infrastructure for development and change (Black et al., 2003; Leung, 2004).

**Classroom Interaction**

One of the key arguments for formative teacher assessment is that, in principle, it is integrated into everyday teaching and learning activities and that it carries an enormous potential for providing useful and helpful teacher feedback, so that students can engage with further learning and/or revise what they have learned. However, feedback, as we will show, is far from a straightforward matter in classroom assessment.

Quite clearly, talk in the classroom, and talk more generally, can be seen as utterances produced by individual speakers as acts of volition. But teachers and students in the classroom do not speak in a social vacuum; they are participants in social interaction. If classroom talk is
construed as part of social interaction between teachers and students and among students themselves, then we need to take a more complexified and dynamic view. The work of Brown (1995), Fulcher (2003), McNamara (1997), McNamara and Roever (2006, chapter 2) and Mohan (1998) has pointed to the need of attending to the co-constructed nature of talk in performance testing of spoken language—the language used by a candidate has to be understood in the context of exchanges between all participants involved. This point also applies generally to formative assessment conducted by the teacher in the classroom, as co-constructed interactions between the assessor and the assessed, whether spoken or written, about language or about other topics. To help us understand this issue more clearly, we can start by looking at classroom interaction in terms of the (teacher) initiation (I)–(student) response (R)–(teacher) evaluation/feedback (E) sequence (Mehan, 1979; van Lier, 1996). Let us consider the following two hypothetical cases:

<table>
<thead>
<tr>
<th>Case A</th>
<th>Case B</th>
</tr>
</thead>
<tbody>
<tr>
<td>I: (Teacher) Where is Britain?</td>
<td>I: (Teacher) Where is Britain?</td>
</tr>
<tr>
<td>R: (Student) To the west of continental</td>
<td>R: (Student) To the west of continental</td>
</tr>
<tr>
<td>Europe.</td>
<td>Europe.</td>
</tr>
<tr>
<td>E: (Teacher) Good.</td>
<td>E: (Teacher) OK, can you tell me where</td>
</tr>
<tr>
<td></td>
<td>you found this information?</td>
</tr>
</tbody>
</table>

The E part in Case A closes the interaction, and the teacher accepts the answer without enquiring further the basis of the student’s knowledge. In Case B the teacher accepts the student response but goes on to explore the basis of this answer. The E part in Case B can lead to teacher–student talk that can show the teacher how (and how well) the student has come to this piece of knowledge (assessment) and what, if anything, the teacher can do to help the student enhance her or his understanding (evaluation/feedback). In this sense the E part in Case B is formative. Of course, in any real classroom the IRE sequence should not be seen as isolated exchanges. It should be understood as part of wider classroom discourse expressing and constituting pedagogic and curriculum practices. (In the next section we will extend this point in a more language teaching focused discussion.)

Black (2001), a proponent of formative teacher assessment working within a Vygotskian tradition, recognises the significance of social interaction and the language discourse associated with it:

One function of assessment would . . . be to help to identify [the zone of proximal development] accurately and to explore progress within it. Thus the dependence of learning on teacher-pupil interaction is a very specific one, linked to the nature of the teacher’s guidance. . . . Most theorists . . . emphasize . . . the importance of language in learning. Interaction takes place through language discourse, which is learned and understood in particular social contexts. . . . It would follow that the nature of our learning depends on the particular “communities of discourse” and its effectiveness on the extent to which its terms and conventions are shared, be it . . . an academic seminar or a school lesson. (pp. 15–16)

From this perspective, talk in the classroom is no longer the property and the product of an individual acting in isolation; it is taken to mean the use of language as discourse in social interaction. Pryor and Torrance (1996) and Torrance and Pryor (1998) show that teachers, when acting as assessors, may be concerned with a number of different matters including opportunistic
use of scaffolding to enhance group learning, use of assessment information for forward planning of the delivery of the curriculum (see next section), and maintaining student motivation. In most educational settings, teachers exercise considerable asymmetric power over students in regulative and pedagogic decision-making in terms of day-to-day classroom construction and representation of the curriculum. This power inequality makes it all the more important to explore teachers’ part in shaping student talk. At the same time, it has been generally recognized that students also bring to curriculum and assessment tasks their understanding and interpretation of what is to be done (often with others) in specific contexts; therefore, teacher assessment has to take account of the agentive aspect in both teacher action and student performance. It follows that there are very good reasons to ask questions about how teachers and students interact and how they construe their interaction through discourse.

**Disciplinary Belief and Classroom Practice**

If we accept that teachers’ evaluation and feedback are key to enhancing student learning, then we need to understand teachers’ judgments on performance. More specifically, we need to attend to teachers’ pedagogic perspectives and disciplinary approaches. In other words, we need to understand how, from teachers’ perspectives, curriculum subjects or topics are constituted, what kinds of teaching and learning are required, and how subject knowledge is translated into pedagogic action. In a discussion on the need to theorize teaching, James (2006) sees assessment in this way:

> Insofar as assessments are intended to assess “something”, that is, some content, account needs to be taken also of the way the subject domain of relevance is structured, the key concepts and/or “big ideas” associated with it, and the methods and processes that characterize practice in the field. (p. 48)

The emphasis on taking account of disciplinary perspectives in studying teacher assessment is important because the relevance and potential contribution of formative assessment to student learning is inextricably related to the ways teachers see the teaching and learning of their subject (for an earlier discussion see Leung, 2004, 2005; Wiliam, 2001). In a fundamental sense, we should be interested in teachers’ epistemological and disciplinary commitments and values in relation to the subjects that they teach. James (2006, pp. 50–52) provides a helpful illustration of this relationship between epistemology, pedagogy, and assessment by presenting (hypothetical) vignettes of assessing the ESL writing of a newly arrived student. These vignettes are situated in the English schooling context where new ESL students are integrated into mainstream classes, including English (as a mainstream subject, not ESL) classes. James provides three vignettes:

- A traditional vocabulary-and-grammar-oriented teacher would set a timed test with items that would cover parsing of parts of speech, punctuating short tests, knowing lists of adjectives and nouns and so on; English language is broadly seen as comprising the components of the language system, e.g. word spelling and spelling rules, classes of vocabulary words and sentence/phrase level grammar rules; knowledge is implicitly acquired through cognitive processes.
- A constructivist teacher would observe the newly arrived student’s participation in English (subject) activities, which would involve reading of a literary text, discussing the literary qualities of the reading text (structure, characterization, etc.), students drafting their own writing with a writing frame, and students sharing their drafts; the teacher
would generate opportunities for the student to engage in on-task classroom talk and help the student to physically manipulate text organization by using text strips; the teacher would also note the new ESL student’s areas of difficulties for further work; knowledge is individually constructed through interaction with other people and texts.

- A teacher with a sociocultural approach would set up the classroom as a writing workshop where all students are to determine their own writing agenda on a common topic or theme; members of the class are to draw on the expertise inside and outside (through published work) the classroom to challenge and develop one another’s work; the dynamic and creative processes involved would produce emergent group criteria for their writing; the new ESL student is encouraged to participate in these activities, and assessment is focused on the student’s participation and the overall outcome of being able to produce writing that would meet with the collective approval of the class; knowledge is developed through participation and apprenticeship.

Ivanič (2004) offers a more detailed account of the teaching of writing which is directly helpful for this discussion. She points out that the teaching of writing, often in the guise of literacy education, has been the subject of constant intellectual and policy debate (in English-speaking countries). The different perspectives and approaches are associated with distinct epistemologies and pedagogies. Ivanič (2004) suggests that there are six approaches (or discourses) to the teaching of writing. The six approaches are summarized as follows:

**Table 1. Teaching Writing Approaches**

<table>
<thead>
<tr>
<th>Approaches</th>
<th>What does writing comprise?</th>
<th>Learning to write</th>
<th>Teaching and assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Skills</td>
<td>Applying rules governing sound-symbol correspondence and syntactic patterns</td>
<td>Learning rules governing sound-symbol correspondence and syntactic patterns</td>
<td>Explicit teaching of rule-governed aspects of written language; focus on accuracy of production/reproduction</td>
</tr>
<tr>
<td>Creativity</td>
<td>Using one’s creative ability to produce text</td>
<td>Learning by writing on topics of interest (to the writer)</td>
<td>Implicit teaching drawing on language experience; focus on interesting content/style</td>
</tr>
<tr>
<td>Process</td>
<td>Executing cognitive and practical steps involved in writing</td>
<td>Learning the processes and the work involved in planning, drafting and revising writing</td>
<td>Explicit teaching and modeling of writing processes; unclear assessment focus, probably some attention to evidence of drafting and revision</td>
</tr>
<tr>
<td>Genre</td>
<td>Producing different types of text for different social purposes</td>
<td>Learning the need for different text types for different social purposes, and the ways language</td>
<td>Explicit teaching of genres; focus on appropriateness of text types and language expressions</td>
</tr>
</tbody>
</table>
### Table

<table>
<thead>
<tr>
<th>Social practice</th>
<th>Participating in communicative/literacy events in ways that involve social relations, uses of technology, social/community goals</th>
<th>Learning to write by writing in real-life (or in simulation) for specific purposes</th>
<th>Explicit teaching where social contexts and purposes are known; implicit teaching where the classroom is constructed to replicate an outside social environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sociopolitical</td>
<td>Knowing that writing is implicated in social and political power and it can also part of the operation of power.</td>
<td>Learning different types of writing and learning to take a position to change or transform existing ‘unjust’ power relations</td>
<td>Explicit teaching of social and political analyses in relation to writing; focus on writing as challenge to existing power relations</td>
</tr>
</tbody>
</table>

**Note.** This table is based on Ivanič (2004).

Ivanič (2004) suggests that these six approaches represent recognizable associations among values, beliefs and practices which lead to particular forms of [teacher] situated action, to particular [teacher] decisions, choices and omissions, as well as to particular wording. (p. 220)

From the point of view of formative assessment, we can see that the kind of work carried out by James and Ivanič, which can be extended to other areas of language teaching and learning, is very useful. It provides insights and analytic frameworks for investigating and understanding our own teaching and assessment practices.

### Professional Development and Professional Infrastructure

Formative assessment is rapidly gaining policy support in a large number of educational systems, for example, Hong Kong (e.g., Education and Manpower Bureau, 2004), Canada (e.g., Western and Northern Canadian Protocol for Collaboration in Education, 2006) and the United Kingdom (e.g., Daugherty Assessment Review Group, 2004; Qualifications and Curriculum Authority, n.d.). Research in this field to date shows that the extent to which the educational value of formative assessment is realized in local contexts crucially depends on how far teachers can put theory into practice in their everyday work (e.g., Black et al., 2003; Davison, 2004; Leung, 2004, 2005; Leung & Lewkowicz, 2006, 2007; Leung & Mohan, 2004; Rea-Dickins, 2000, 2001, 2006). In order to exploit the educational potentials of formative assessment to the maximum, language (and other subjects) teachers and policy makers should ask the following questions in relation to their local working context:
• Is the education system sufficiently attuned to the potential and actual contribution of formative assessment to the quality of learning?
• Is formative assessment given official recognition within the education system (e.g., is there reporting of teacher formative assessment in the profiling of student overall achievement)?
• Is there professional support for teachers to develop a reflexive understanding of their classroom practice and their disciplinary beliefs as individuals and as a collectivity?
• Are there opportunities for teachers to develop alternative and/or additional disciplinary knowledge and understanding to enable them to adopt and adapt their assessment practices?

The answers to these questions will help formulate strategies required to push the formative assessment agenda forward.
References


Should Authenticity Continue to Trouble Language Testers (and Teachers)?

Jo Lewkowicz  
Institute of Applied Linguistics  
Warsaw University  
Warsaw, Poland

Introduction

In a review article on recent developments in language testing and assessment, Leung and I (Leung & Lewkowicz, 2006) argued that perhaps the single most pervasive issue in the field has been the question of authenticity. Test authenticity has been widely discussed not only in its own right (e.g., Bachman, 2000; Bachman & Palmer, 1996; Lewkowicz, 1997, 2000; Wu & Stansfield, 2001), but also in discussions related to English for specific purposes (ESP) testing (e.g., Douglas, 2000), alternative assessment (e.g., Brown & Hudson, 1998), washback (e.g., Watanabe, 2004), and validity and reliability (e.g., Messick, 1994), among others. One of the reviewers of the paper asked: so what? The reviewer, I believe, was not questioning the fact that authenticity has been a key consideration in the language testing debate over the past 15 or so years, but rather was asking whether anything new could be added to the discussion. The underlying suggestion seems to have been that there is a common understanding of authenticity which can readily be operationalized in a way that makes tests immediately discernible as authentic (or not). To explore this issue, I address the following three questions:

• Why is authenticity an important consideration in language testing and assessment?
• What is it that makes a test/assessment authentic?
• What are the difficulties that impede the development of (maximally) authentic tests/assessments?

Why Consider Authenticity?

I start with the issue of why in order to establish that authenticity is an important consideration for the development of tests/assessments. Only then does it make sense to go on to determine a common understanding of what is meant by authenticity in testing and assessment and how to operationalize it in test/assessment situations.

There are allegedly three reasons for wanting to make tests authentic. The first reason relates to generalizability from test to nontest situations. Testing is concerned with eliciting a sufficient sample of language from test takers to be able to make generalizations from their performance on a test to their likely performance in target language use situations. If, for example, the test is to determine whether the candidates have sufficient language to survive an academic programme abroad, then the test scores should provide information about which candidates have the necessary language for academic study and which do not. In other words, there needs to be a close synergy between the language tested and the language that is needed for study abroad.
The second reason for considering the language to be tested in terms of test to nontest situation is that of washback, that is, the effect a test will have on teaching and learning. Washback is a complex notion which may relate not only to the effect a test has on the content of teaching, but also to the textbooks chosen for teaching, classroom organization used, or technology employed in teaching (see Alderson & Wall, 1993, and Bailey, 1996, for a full discussion). Yet, in high-stakes tests, in particular, the test will likely have an impact on what is taught. Students expect teachers to cover what will be tested and to provide practice in test tasks. If what is to be taught is to be useful for target language use, then what is tested also needs to reflect the language test takers will need outside the test and classroom situations.

The third reason for considering the language to be tested is that of test-taker motivation. It is purported that tests that are authentic are likely to be perceived as relevant and thus to influence test takers’ performance. Bachman and Palmer (1996, p. 24) suggest that

One way in which test takers and test users tend to react to language tests is in terms of the perceived relevance, to the TLU domain of the test’s topical content and the types of tasks. It is this relevance, as perceived by the test taker, that we believe helps promote a positive affective response to the test task and can thus help test takers perform at their best.

Yet there is no real evidence to support this claim. My own research done some years ago (Lewkowicz, 1997) suggests that test takers are more pragmatic: they are primarily concerned with doing well and having had practice in the types of tasks used in the test rather than in whether the test is authentic or not. And, to my knowledge, there have been no studies in this area contesting my findings.

What Makes a Test Authentic or Not?

Regardless of whether test takers pay attention to the relationship between a test and nontest language use, it appears from the above that what is tested needs to relate to the language that learners are likely to need in the real world outside the classroom if we are to be able to make generalizations from the test to nontest language performance. And indeed, it is this relationship between test task and target language use task that is seen as an indication of test authenticity. Authenticity is thus defined as “the degree of correspondence of the characteristics of a given language test task to the features of a TLU [target language use] task” (Bachman & Palmer, 1996, p. 23).

In order to operationalize this definition and determine the level of test authenticity, one needs to refer to the framework of language task characteristics proposed by Bachman and Palmer (1996). This framework describes five aspects of tasks which need to be considered, namely, task setting, test rubric, input, expected response, and the relationship between input and the response. In its entirety, the framework is very detailed (see Bachman and Palmer, 1996, p. 49-50). What is immediately evident from this framework, however, is that the test task is seen as being made up of both input (e.g., a reading passage in a test of reading comprehension, and test questions) as well as the response (e.g., the underlining of the correct response in a multiple choice test or a short written response to short answer questions), and both input and response may vary in terms
of authenticity. Thus, in any test or, indeed, test task, determining a single level of authenticity would appear very difficult, if not impossible. Authenticity seen in this way is not regarded as an absolute quality of a test, either present or absent, as was previously considered, but rather a relative quality, allowing for a test, or at least the separate components of a test, to be more or less authentic.

This conceptualization of authenticity has two major advantages. It recognizes that one needs to look beyond test input (i.e., text and tasks) and acknowledges that what test takers have to do with the language is equally as important (cf. with early definitions of authenticity such as those by Morrow, 1979, and Porter & Roberts, 1981, where the input was seen as the defining issue in determining authenticity). Thus, it is recognized in this definition of authenticity that has been accepted since the late 1990s, that a multiple choice reading comprehension test based on authentic (nonpedagogic) texts cannot be regarded as authentic because TLU situations rarely if ever require ticking the correct answer from three or four options that have been provided: at best it may be regarded as partially authentic, with the texts contributing to task authenticity.

This conceptualization of authenticity also allows us to acknowledge that TLU situations will vary for different learners and what test takers are able to relate to will also vary. This would appear to be in line with Widdowson’s (1978, 1990) notion of the need for authentication, where he argues that it is not sufficient to take texts from real-world (nonpedagogic) contexts for teaching (and testing) purposes, as not all second language/foreign language (L2/FL) learners will be able to relate to, that is authenticate, every text. (It is worth noting here, that this is also true for native speakers of a language who are readily able to relate to and fully comprehend some texts but not others.)

A further advantage of Bachman and Palmer’s (1996) conceptualization of authenticity becomes apparent when it is recognized that authenticity is seen as only one of six equally important qualities of a test determining any test’s usefulness. Test usefulness is seen as ultimately the most important overarching quality of any test, and to achieve maximum usefulness one needs to balance

Reliability + Construct Validity + Authenticity + Interactiveness + Impact + Practicality

Given that tests are administered for very different reasons, it may not always be necessary or desirable to aim at maximizing authenticity. In a low-stakes achievement test, for example, construct validity or practicality may become the primary test qualities to be enhanced.

An issue worth pondering, however, is why interactiveness is seen as a separate test quality and not part of and related to authenticity. Interactiveness is defined as the “extent and type of involvement of the test taker’s individual characteristics in accomplishing a test task” (Bachman & Palmer, 1996, p. 25). A task may require a high degree of test-taker interaction such as in the case of a one-to-one interview or a simulation where, for example, the test taker is required to play the part of a nurse talking to a hospital patient’s relative (played by the examiner). Alternatively, the task may require limited interaction such as when word processing a hand-written text. Yet what is relevant here in terms of being able to generalize from test to nontest situations is not simply the level of interactiveness required, but whether it is the type of task the
test taker could reasonably be required to perform in nontest TLU situations. The level of task interactiveness thus must, in part, determine authenticity. Therefore, it would appear to be impossible to separate out these test qualities; rather, they should be seen as two aspects of the same quality—authenticity—which is, in part, dependent on the interactiveness of the task. This is, in fact, the way authenticity was originally perceived by Bachman (1990) when he distinguished between situational authenticity and interactional authenticity. The change in conceptualization limiting authenticity to what was previously labeled situational authenticity came about because of the realization that only test tasks can vary in their level of authenticity, while real-life tasks as well as test tasks may differ in terms of interactiveness (L. Bachman, personal communication, October, 1999). This more restricted view of authenticity seems now to have become the accepted canon in testing circles.

It should be apparent from this discussion so far that although relating test task to TLU tasks is a paramount consideration when developing language tests, how this should be done remains problematic because we have not yet arrived at a single, unquestionable understanding of authenticity as a concept.

**Operationalizing Authenticity**

I now turn to the issue of whether and to what extent we are able to operationalize this nebulous concept to ensure maximally authentic (for their particular purpose) tests. As indicated above, Bachman and Palmer (1996) provide a comprehensive characterization of tasks to help estimate the level of authenticity of any test task in relation to TLU tasks. This characterization, however, is fairly complex, and although it has been employed in a number of research projects, it does not appear to have been universally adopted when developing tests, even though sufficient time has elapsed for test developers to be familiar with the framework of task characteristics. The framework’s complexity, even without taking into account task interactiveness, seems to militate against universal adoption. Thus, there seems to be a continuing gap between research and theory, on one hand, and test development practice, on the other.

The framework’s complexity, however, is not the only problem facing test developers who want to ensure maximally authentic tests. The underlying premise of the current definition of authenticity is that TLU tasks and situations can be accurately identified and described, thus allowing test developers to match test task characteristics to the TLU domain. This may be at least partially true of English for specific purposes/language for specific purposes (ESP/LSP) tests, where test takers are relatively homogeneous. Indeed, Douglas (2000, p. 2) identifies authenticity as one of the distinguishing characteristics of an LSP test (the other one being “the interaction between language knowledge and specific purpose content knowledge”). However, the less homogeneous test takers are, the more difficult it is to predict their future language needs and the situations in which they will use the language they are learning. It may be very difficult if not impossible to identify, for example, the TLU of adolescents learning English as a foreign language in Ukrainian secondary schools. Some may want to travel to English-speaking countries and may need to communicate with native speakers of the language; others may travel to Europe and use English as a lingua franca, while yet others may remain in Ukraine and have limited exposure to English in their studies or working life. Yet these students would likely have similar interests and topical and cultural knowledge, facilitating to some extent the selection of
TESOL Symposium on English Language Assessment – Jo Lewkowicz

test input they would be able to relate to. Where test takers come from a much more diverse background, for example, for the International English Language Testing System (IELTS) and the Test of English as a Foreign Language (TOEFL), identifying input that all may relate to becomes increasingly difficult.

The issue of characterizing TLU does not limit itself only to the future language needs of test takers in terms of what skills they will need or what functions they will be required to perform. It also raises the question of what variety of language should be described and tested (and hence taught). Standard varieties, such as Standard British or American English, are the most accurately described to date and hence the most commonly adopted for testing purposes. However, these may not reflect the TLU domain of all test takers, i.e., they may not be the most authentic. For a large number of Poles seeking employment in Ireland, for example, being acquainted with Standard British English and being able to show a level of language proficiency attained on one of the Cambridge suite of tests, e.g., First Certificate of English (FCE) or Certificate of Proficiency in English (CPE), may not reflect their true ability to communicate with Irish users of English. (Many report considerable difficulty with understanding their Irish counterparts at work.) So the growing use of English as a lingua franca is posing considerable challenges to determining TLU accurately, particularly in situations of large-scale testing.

One further problem of operationalizing authenticity in terms of relating test tasks to TLU tasks needs to be considered. When developing tests, it is up to the test developers to ascertain a desired level of authenticity and to ensure that this level is apparent to test takers in the test produced. There may, however, be a difference in perception between test developers and test takers and even among the test takers themselves. In addition, it must be recognized that test development is a reiterative process often involving moderation of test tasks. Such moderation may lead not only to changes in the task that were not anticipated at the test specification stage, but also to changes that affect the degree of correspondence between test task and TLU tasks (i.e., in the degree of authenticity). In one of the studies I carried out for my PhD, a test that was initially to be highly authentic in terms of test input, test tasks, and the written outcome to be produced by the test takers, turned out to be well below the level of authenticity specified at the initial stages of test development. The lowering of authenticity came about as a result of moderation, and worth noting here is that the moderators made very few appeals to authenticity. Their primary concerns were issues of practicality and ensuring the input could stand alone in the test situation and the test takers would understand the language used. Trying to fix authenticity at the test specification stage may therefore be unrealistic.

The Future of Authenticity

There is no doubt that authenticity as discussed above is a highly complex notion that is difficult to implement in test situations. The question thus arises: should we abandon our quest for authenticity? My answer is both yes and no for the following reasons.

Test development needs to focus on ensuring that tests measure as accurately as possible language ability, and if the desired ability is to communicate in the language, we need to ensure that we accurately test this construct. In other words, we need ensure the construct validity of the test. That construct validity will in part be determined by how our test tasks relate to real-life
tasks and real-life language use, and in this respect, despite its complexities, I believe we cannot abandon authenticity.

Authenticity is, however, a nebulous concept, perhaps unattainable in any objective way in practice. Yet what we seem to be aspiring for is some kind of scientific formula that allows us to estimate a level without having an appropriate scale for doing so. Authenticity relates to qualities of human interaction and social practices which are dynamic in nature. By trying to pinpoint it with scientific accuracy, we may be pursuing a chimera. From this perspective, therefore, I think the pursuit of authenticity could be abandoned.

As a term, *authenticity* seems to lead to unnecessary dispute and confusion. Gilmore (2007), in fact, reminds us that it is generally associated with something desirable and contrasted with unreal/inauthentic language or tasks which are undesirable. A more neutral description such as *experiential relevance* would possibly help to avoid this problem. The question ultimately is whether the tasks we employ in a test situation allow us to determine language ability and are relevant to test takers in terms of what they have been taught and, in turn, to language as it is used in the real world.

I am not suggesting we adopt a new term for a well-established concept—pouring old wine in to a new bottle and relabeling it. I believe we need a term which will be less emotive and will ensure that we relate test tasks and assessments to the needs of specific test takers, addressing the question *relevant for whom*? We also need to recognize that in determining test content we cannot view the test in isolation of teaching. Most tests come at the end of a period of formal instruction, and therefore it is not simply that what is tested will impact on what is taught: what is taught will also be tested. We need to bring this relationship into the equation, which is difficult if we continue to use the term *authenticity* to refer to both the language and materials used in teaching and the tasks used in testing. Test takers are likely to be able to comment on whether they perceive test tasks as relevant to their needs and to what they have been taught more so than whether they think the task authentic. This does seem to be compatible with Bachman and Palmer’s (1996) reasoning outlined above, namely, that perceived relevance of tasks is likely to positively affect test taker’s performance, though I would argue here that relevance includes more than situational authenticity.

Whether tasks are authentic may, thus, not be as crucial as has previously been claimed, but whether tasks are perceived as relevant by test takers who have been through a period of learning, by test developers who understand the aims of the test/assessment and by other stakeholders interested in understanding reported test taker performance, would appear to be indisputable.
References


Closing Session: Summary

Olga Bessonova
Donetsk National University
Donetsk, Ukraine

Dear Ladies and Gentlemen:

First, let me express my cordial thanks to TESOL and TESOL Ukraine for inviting me to attend the symposium, for the opportunity to summarize the key points made by today’s speakers, and to share with you some thoughts and ideas about the problems raised at the Symposium on English Language Assessment.

Assessment is an umbrella term, which includes diverse practices and approaches. The specific focus of the symposium on the effective ways of developing English language tests and assessment procedures in line with national and international standards and practices is of great importance. This approach promotes

- further understanding of fundamental issues and concerns involved in the appropriate use of language tests
- understanding of conceptual bases of language test design and development
- the ability to critically perceive new developments in language testing in order to make informed decisions.

This symposium is a very important event for Ukraine because this country is implementing an education reform. In January 2004, Ukraine introduced a new National ELT Curriculum for secondary schools, having specified students’ learning attainments at A1 level for primary school, A2+ for secondary school, and B1+ for high school. The most recent development is the endorsement of a new National ESP Curriculum for Universities, which identifies the attainment level for a bachelor’s degree as B2. The Independent Testing Centre, authorized by the government to provide professional examination services to school leavers, has identified English examination for university applicants as one of its priorities.

The British Council Ukraine, together with the Ministry of Education Ukraine and the National Testing Centre, has launched a new project to support Ukraine’s standardised testing initiative. The aim of the project is to set up a group of professional test-item writers who will be able to link assessment to the Common European Framework of Reference for Languages (CEF), construct good language proficiency tests, and participate in large-scale testing programmes. The project team is being trained by an international team of experts from Scotland, Hungary, and Norway. The project is working toward professional content, transparent and practical exam procedures, with a view to relating its results to the CEF.

As a language proficiency measurement tool, testing affects all aspects of language education. Teaching methods and learning strategies, the overall efficiency of language learning, as well as the content of language courses are all affected by the nature and purpose of testing in the assessment system. Testing contributes to the system of assessment in such a way that students’
performance improves. Thus, testing may be regarded not only as a measurement instrument but also as an improvement tool.

The efficiency of language education can be improved in various ways; for instance, education officials can use test scores to target educational resources to low achieving schools or geographic areas; testing can be used to shape and pull teaching methodologies in desirable ways; testing can be used to motivate teachers to improve their teaching; testing gives teachers information with which they can target remediation; national assessment can support cross-national comparisons which can lead governments to commit a larger share of the national budget to education. All these things are very important for Ukraine’s English language educators as well as for the public at large. The attitude toward testing is perfectly expressed by the title of Alderson’s (1999) plenary address to the Third Annual Conference on Current Trends in English Language Testing in Dubai, “Testing is Too Important to Be Left to Testers.”

In spite of Alderson’s global message, testing is the work of specialists with its theoretical background and practical application. It is very important that today’s symposium has offered the opportunity of plunging, with the speakers’ assistance, into the theory and practice of testing.

Our colleagues have discussed today such key issues of assessment as transparency, formative assessment, and authenticity. The choice of the issues meets the needs of the participants and makes it easier to communicate across languages, educational sectors, and national boundaries. Let me remind you of the most important points of the extremely interesting morning presentations and the lively afternoon workshops.

**The Importance of Being Transparent: Benefits to Society, Schools and Students of Using a Common Language Assessment Framework (Jeff Connor-Linton, Georgetown University)**

Jeff Connor-Linton looked at the notion of transparency, discussing the values of transparency at different levels of application, that is, the way it benefits language testing in international, national, institutional, and pedagogical contexts. This approach allows a very broad interpretation of transparency. It implies not only openness and accountability; it also enhances the face validity of a test, encourages learners to internalize the standards against which their performances are judged, and it creates a positive washback effect.

Connor-Linton maintains that at the international level, transparency can be attained by adopting shared, explicit standards of language performance. Being transparent is definitely beneficial for a country. If testing practices in a particular educational system are transparent, the marketplace understands the value of its graduates, because employers, academic institutions, teachers, and students share a common understanding of what a particular rating on the test means. Besides, shared performance standards serve as a common framework, obviating the need to reinvent a language assessment system and enabling more efficient use of limited language education resources.

Jeff Connor-Linton focused on the CEF, summarizing its main features, because it serves as a foundation for his further arguments. A closer look at the CEF provides evidence that the document can play an important role in language education reform in Europe and beyond. The
main strength of the CEF in language policy making is its multipurpose nature: The CEF offers the basis for syllabus design, discusses various types of assessment, promotes learning autonomy, and other purposes. Because of its multifunctional nature, the CEF can help to shape objectives and priorities of language education, teaching and learning methods, materials design, and assessment techniques. The CEF describes in a comprehensive way what language learners have to learn in order to use a language for communication and what knowledge and skills they have to develop so as to be able to act effectively. The description also covers the cultural context in which language is set. The Framework defines levels of proficiency which allow learners’ progress to be measured at each stage of learning and on a life-long basis. (Council of Europe, 2001, p. 1)

Connor-Linton stressed that, with its six-level scale, the CEF provides a way of describing diversity and at the same time facilitates relations between different systems of qualifications across different countries. In his afternoon workshop session, Connor-Linton brilliantly demonstrated the practical application of his ideas training the audience to use rating rubrics as a teaching tool. He encouraged participants to approach different types of rating scales and CEF descriptors analytically and interactively. He led the participants, gradually, from Level A1 (Step A1) to Level C2 (Step C2) of his master class.

What are the benefits of transparency at national level? Connor-Linton mentioned three benefits. The first benefit is that the transparency of the criteria for evaluating language helps ensure accountability of test developers and administrators and therefore the fairness of tests. The second benefit deals with promoting equal access to educational resources, thus enhancing the reputation of the educational system and the government. The third benefit is that transparency allows schools to rely on ratings of students’ language abilities and place them more accurately in their curricula (saving educational costs and improving student motivation and performance).

Connor-Linton strongly supports his arguments referring to the Code of Fair Testing Practices, developed by the Joint Committee on Testing Practices of the American Psychological Association. The code incorporates regulations on developing and selecting appropriate tests, reporting and interpreting test results, administering and scoring tests, and informing test takers. Such detailed and transparent descriptions meet the standards of fair testing. The CEF also provides transparent criteria enabling the framework to describe different qualifications and to define different learning goals. One of the strong points of the document is its openness and flexibility. Extrapolating the CEF principles to modern languages teaching and learning and to other ethno-specific studies is very important and should be encouraged. In this sense, the CEF principles should be regarded as valuable experience, on which any national system of language education may be grounded. Another valuable document designed to allow test developers to relate tests based on a range of testing approaches to a common set of performance standards is the Manual for Relating Language Examinations to the Common European Framework of Reference.

At the institutional level, the principle of transparency in language assessment can facilitate language curriculum development and help make language curricula a vibrant resource for language teachers. Elaborating on this idea, Connor-Linton stressed that transparency enables teachers to rely on the language ratings of new students in their classes and more efficiently
respond to those students’ individual needs. Students can use performance descriptors of a particular test as a rubric for self-assessment and a tool for more effective language learning inside and outside the classroom. He pointed out that characterizing levels of development from several perspectives allows the teacher or tester to “triangulate” the different perspectives in order to reach a more reliable and useful assessment of the individual learner. Triangulating different elements of curriculum (levels of development; language skills; qualitative aspects of language use; and genres, domains, and topics) is a valuable resource and guide for teachers’ syllabus design and class planning.

Why is transparency tremendously important at classroom level? The answer that Connor-Linton gives is predictable: because making language evaluation criteria transparent to learners can help them participate more actively in their own learning. He stressed that transparency helps them develop strategies for improving their own communicative skills and gives them more control over their own second language development. Students can use test performance descriptors as a rubric for self-assessment and a tool for more effective language learning inside and outside the classroom. Transparency at classroom level helps students internalize the goals of instruction and apply them to their own language learning.

**Formative Assessment: Time of Learning? (Constant Leung, King’s College London)**

The presentation by Constant Leung dealt with the key principles of formative assessment with particular reference to second and foreign language education. “Formative assessment is an ongoing process of gathering information on the extent of learning, on strengths and weaknesses, which the teacher can feed back into their course planning and the actual feedback they give learners” (Common European Framework of Reference for Languages: Learning, Teaching, Assessment, 2001). As we can judge, formative assessment is designed to identify learners’ strengths and weaknesses to effect remedial action. The results of formative assessment feed back into instruction. As Leung pointed out, assessment has played a large part in education initiatives and reforms designed to improve student attainment in countries such as Australia, the United States and the United Kingdom during the past 15 years. In his talk, he focused, on the one hand, on the pedagogic usefulness of classroom-based formative assessment in promoting learning and educational achievement and, on the other hand, on some theoretical and practice-related issues that require further elaboration.

Providing context for the discussion, Leung resorted to recent experience in England, where in 1997 the government set an attainment target for 11-year-olds in state-funded schools: 80% of the students in this age group were expected to achieve Level 4 in English by 2002. Regular (formal standardized) testing of student performance was used as a policy instrument to push up achievement. The statistics, which Leung provided, clearly show that there have been gains in the percentage of students achieving Level 4 in the past 10 years. Still, the official targets have not been met because the high-stakes nature of formal standardized testing has led to a narrowing of the English curriculum and “teaching to the test.” In response, the British government has started to encourage teachers to adopt formative assessment, also known as *assessment for learning* in their routine classroom practice. There is no doubt that feedback on the effectiveness of student learning is generally of interest to teachers as well as students. Continuous feedback to both teacher and learner is important for “making decisions regarding appropriate modifications
in the instructional procedures and learning activities” (Nitko, 1988). Formative assessment enhances student learning and achievement, but feedback works only if “the recipient is in a position (a) to notice . . . , (b) to receive . . . , (c) to interpret . . . , (d) to integrate the information” (Common European Framework of Reference for Languages: Learning, Teaching, Assessment, 2001).

Leung emphasized a range of teacher actions that included using exploratory questioning and increasing wait time for student answers to generate opportunities for sustained discussion, providing helpful feedback designed to improve student learning and performance, and using the information produced by summative tests for formative purposes. Leung is sure that such an approach can be transferred to other subjects. However, for subject specialists to carry out assessment for learning effectively, it is important for teachers to have an informed understanding of the key notions, such as exploratory questioning and helpful feedback, with reference to their subject disciplines.

Leung highlighted a number of issues that are important for English language teachers. Among them are articulation and clarification of disciplinary beliefs and classroom practices at individual and collective levels, attention to the co-constructed nature of classroom interaction and its implications for assessment of learning outcomes in the English classroom, and professional infrastructure for development and change. Leung supported his arguments using classroom data and analyzing think-aloud protocols. He also addressed conceptual issues such as formative direction and professional issues such as teacher preparation.

Should Authenticity Continue to Trouble Language Testers (and Teachers)?
(Jo Lewkowicz, Institute of Applied Linguistics, University of Warsaw)

The third presentation considered authenticity, another guiding principle that governs good test design, development, and analysis. Alongside transparency, validity, reliability, practicality, washback, security, and usefulness, “authenticity is at ‘the heart of the matter’ of language testing” (Bachman, 1990). That is why, as Lewkowicz mentioned, the issue of authenticity in language testing has been extensively discussed and our understanding of this notion has developed considerably. However, the question is whether we have a common understanding of authenticity and whether it can be operationalized in a way that makes tests immediately discernable as authentic. Exploring the issue, Lewkowicz addressed the importance of the principle of authenticity, those characteristics which make a test authentic, and difficulties that impede operationalization of authenticity.

Answering the question “Why consider authenticity?” Lewkowicz speaks about three reasons for wanting to make tests authentic. The first reason relates to generalizability from the test to non-test situations. In other words, the language tested (the candidate’s performance on a test) needs to be similar to the language needed in a real life situation (the candidate’s likely performance in target-language-use [TLU] situations). Another reason for considering the language to be tested in terms generalizability is washback, that is, the effect a test will have on teaching and learning. Lewkowicz stressed that the washback effect is especially important in high stakes tests. The third reason for considering the language to be tested is test-taker motivation. Tests that are claimed to be authentic are likely to be perceived as relevant and thus
to influence test-takers performance. Yet, as Lewkowicz points out, there is no real evidence to support this claim. Her own research suggests that test takers are more pragmatic: They are primarily concerned with doing well on the test, and they do not care whether the test is authentic or not.

The indication of test authenticity is the relationship between test task and TLU task. Lewkowicz refers to Bachman and Palmer’s (1996) concept of authenticity. They defined it as “the degree of correspondence of the characteristics of a given language test task to the features of a TLU task” (p. 23). To operationalize this definition and determine the level of test authenticity, Lewkowicz recommended analyzing Bachman and Palmer’s framework of language task characteristics. This framework describes five aspects of tasks that need to be considered: task setting, test rubric, input, expected response, and the relationship between input and response. Authenticity seen in this way is regarded as a relative quality, allowing for a test, or at least the separate components of a test, to be more or less authentic. Such an approach to authenticity has two major advantages. It recognizes that one needs to look beyond test input (i.e., text and tasks) and acknowledge that what test takers have to do with the language is equally as important. In line with this conceptualization, we acknowledge that TLU situations will vary for different learners, as will what test takers are able to relate to. Besides, it is recognized that authenticity is only one of six equally important qualities of a test determining any test’s usefulness. Usefulness is seen as ultimately the most important quality of any test. To achieve maximum usefulness one needs to balance

Reliability + Construct Validity + Authenticity + Interactiveness + Impact + Practicality.

On the other hand, it has been pointed out that in constructing low-stakes achievement tests, validity or practicality, and not authenticity, comes first.

Lewkowicz concluded that although relating test task to TLU tasks is rightly a paramount consideration when developing language tests, how this should be done remains problematic because there is no single, unquestionable understanding of authenticity as a concept. She noted that test developers who want to ensure test authenticity face a number of problems, the framework complexity among them. Besides, test takers are not always a homogeneous group. The less homogeneous test takers are, the more difficult it is to predict their future language needs and the situations in which they will use the language they are learning. Lewkowicz also mentioned the problem of language varieties, because the growing use of English as a lingua franca is posing considerable challenges to determining TLU accurately, particularly in situations of large-scale testing.

The degree of authenticity also depends on one more problem, which the speaker highlighted. When developing tests, the test developers have to ascertain the desired level of authenticity and to ensure that this level is apparent to test takers in the test produced. Perceptions may differ between test developers and test takers, and even among the test takers themselves. The lowering of authenticity may come about because of moderation. Trying to fix authenticity at the test specification stage may therefore be unrealistic.

Lewkowicz drew the conclusion that authenticity is a highly complex notion that is difficult to implement in test situations. The natural question arises: Should the quest for authenticity be
abandoned? In her answer to this question, Lewkowicz showed that she is of two minds. On the one hand, to ensure construct validity of the test, one cannot abandon authenticity altogether. On the other hand, authenticity is a vague concept, perhaps unattainable in any objective way in practice. Because pinpointing authenticity with scientific accuracy is impossible, it could be abandoned, but Lewkowicz posited a possible balance in a more neutral description such as test relevance, or experiential relevance, which would possibly help us avoid this problem.

In Lewkowicz’s afternoon workshop session, the discussion centred on the practical application of authenticity. Participants discussed whether the texts suggested for analysis were authentic, modified, or written for pedagogic purposes, and analyzed to what extent the framework helped determine the authenticity of textual input.

**Conclusion**

The speakers insights into the problems discussed were very detailed and profound, exposing new approaches to language assessment. I am sure that the participants appreciated the speakers’ help in learning more about the field of assessment.

The assessment symposium will stimulate cooperation between language teachers, students, publishers, and specialists in language testing; and it will help us to overcome language barriers and boundaries of national language systems. The independent testing reform in Ukraine may influence the further development of language education, which will lead to the reconsideration of aims, needs, and content; to the development of the harmonious assessment system; and to the establishment of closer links between the secondary and higher education.

Thank you very much for your attention.
References

Alderson C. (1999, May). *Testing is too important to be left to testers*. Plenary Address to the Third Annual Conference on Current Trends in English Language Testing, United Arab Emirates University, Al Ain and Zayed University, Dubai Campus.


